



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Scaling Trust and Safety Decision Systems Across Billions of Content Items

Bharath Kandati

Independent Researcher, India. E-mail: bharathkandatireddy@gmail.com

Abstract

A large-scale music streaming platform operating a catalogue of more than 250 million tracks, with tens of thousands of new tracks arriving daily through distributor pipelines, cannot review its content by hand. At one minute per track, a single pass over the catalogue would consume centuries of person-time. Machines therefore make the decisions, and human judgement is reserved for a narrow uncertain band. This paper is a practitioner account of the machine decision system that resulted. It argues that the hard problems at this volume are not classification problems but governance problems: where a threshold sits, who owns it, which decisions may be automated, what counts as evidence, and what happens when the institution has not decided what a category means. Six design commitments carried the system. Harm is treated asymmetrically: precision is a promise to individuals, recall a promise to the ecosystem, and the false-positive bound is expressed as a count of artists wrongly flagged per month rather than as a rate. Review capacity is allocated on two axes, model uncertainty and stakes, and review is conducted per campaign rather than per item. The historical backfill is separated from the live path by a hard capacity reservation and sequenced by economic urgency rather than upload date. Detection, classification, and enforcement are kept as separate layers, so that the model measures and the policy decides, and correctness is estimated through converging evidence rather than a single oracle. The account is retrospective, no controlled evaluation is reported, and no operational metric is released; each place where a number would ordinarily stand is marked.

Keyword: trust and safety; content moderation at scale; human review allocation; asymmetric harm; AI-generated music detection; platform governance

I. Introduction

Trust and safety engineering at platform scale is usually described as a detection problem. A model is trained, a threshold is set, content is scored, and enforcement follows. That description survives contact with a small catalogue and fails on a large one. On a catalogue of more than 250 million tracks, with tens of thousands of new tracks ingested daily, the detection model is the part of the system that is least in doubt. What is in doubt is everything around it: how a threshold gets chosen and who is accountable for the choice, which of the resulting decisions a machine is permitted to make alone, how the scarce human judgement in the system is spent, how the past gets re-decided without delaying the present, and how anybody knows whether the whole apparatus is right when the question it answers is partly a definitional one.

This paper reports how those questions were answered in practice. The author led engineering for AI trust and safety at a large-scale music streaming platform, where the team built AI music detection across the catalogue described above. The account is deliberately architectural and operational rather than empirical. It does not present a new model, a benchmark, or a controlled comparison. It presents the decision system: its topology, the rules that governed routing, the constraints that were treated as inviolable, the tradeoffs that were made on purpose, and the failure modes that only become visible at scale.

Two facts set the problem. The first is arithmetic. Human review is not an option at this volume, and not because of budget. At one minute per track, a single pass over the catalogue would consume centuries of person-time. Machines make the decisions; humans get the uncertain band. Every design choice downstream is a consequence of that sentence.

The second fact is economic. Royalties are a zero-sum pool. Every stream on an undetected AI-generated track dilutes payouts to real artists. The decision therefore has a deadline, and the deadline is not a service-level target invented by the engineering team. It is the moment money moves. Once payouts have been computed on a stream that should not have counted, remediation is expensive and incomplete. A detection system that is accurate but late is, in the part of the problem that matters most, a system that did not work.

Those two facts produce a third, which organises the rest of the paper. The two failure modes are asymmetric, and the asymmetry runs deep enough that it should be treated as a design axis rather than as an evaluation

detail. A false positive, flagging a real artist as AI, is a concentrated harm to one identifiable person: demonetisation, reputational damage, backlash, legal exposure. A false negative is a diffuse harm with no single victim: an eroded royalty pool, degraded recommendations, an open door to industrial-scale fraud, invisible until it is systemic. Precision is a promise to individuals. Recall is a promise to the ecosystem. Where the threshold sits between them is a values statement, not a tuning parameter.

1.1 Contributions

This paper contributes the following, all drawn from operating the system rather than from simulation.

1. **A harm-first threshold discipline.** The false-positive bound is treated as a hard shipping gate and expressed to stakeholders as a count of artists wrongly flagged per month rather than as a rate, because rates hide what thresholds decide. Confidence-band boundaries are derived from that bound. When the derived human band exceeds review capacity, capacity is what flexes.

2. **A three-lane routing architecture** in which the confident majority in both directions never reaches a human, the uncertain middle is the only slice where judgement is spent, and enforcement is proportional to confidence and harm with every decision attributable to the signals that fired and the threshold that applied.

3. **Two-axis review allocation.** Uncertainty is the first cut and the only reason the arithmetic works, but stakes are the second cut: money moving, audience exposure, and the identifiability of who gets hurt if the decision is wrong. Appeals are a standing exception that jumps every queue.

4. **Cluster review as a first-class primitive.** Linked items collapse into one case before a reviewer sees them. Per-item review is how a team drowns; per-campaign review is how it keeps up with an adversary who operates in campaigns.

5. **A backfill discipline for re-deciding the past under a live guarantee**, combining tiered scoring, strict lane separation with asymmetric priority, and sequencing by economic urgency rather than upload date. At this volume, latency is not an infrastructure property. It is a prioritisation property.

6. **A correctness framework for contested ground truth**, built on separating detection from classification from enforcement, and on four independent instruments that make miscalibration visible. High reviewer disagreement is routed to policy rather than to more training data, because a category where reviewers disagree with each other is the policy announcing that it has a hole.

1.2 What this paper does not claim

The underlying operational data is confidential, so this paper releases no operational metric. Two figures about the platform are given, a catalogue of more than 250 million tracks and tens of thousands of new tracks ingested daily. Where a structural claim would normally be accompanied by a number, the claim stands alone and the gap is marked; Section 10 lists in full what is not released. This is a deliberate choice rather than an omission. A system whose errors become visible quickly in analysable patterns is a stronger thing to be able to demonstrate than a decimal nobody can substantiate.

1.3 How this account was produced

This is a retrospective practitioner account, and its provenance bounds what it can be. It was produced from the author's own structured written recollection of a programme he led. No internal document, dashboard, or dataset was consulted in writing it, no colleague was surveyed, and no measurement was performed for it. The subject is a single programme at a single platform, in one content domain, music audio, covering the period in which detection was scaled across the catalogue and the historical backfill was completed. Nothing quantitative was released for publication.

The reader should therefore treat the operational claims in Sections 4 to 9 as a practitioner's account of how a system was designed and run, not as an audited record of how it performed. What the paper offers as evidence is the design rationale, the ordering rules, and the constraints that were treated as inviolable. Those are checkable as argument even where the outcomes are not checkable as data.

2. Related Work

Algorithmic moderation and platform governance. The automation of content moderation has been analysed as a socio-technical and political problem as much as a technical one. Gorwa, Binns, and Katzenbach [1] set out the technical and political challenges of automated platform governance, including the opacity of classifier-driven enforcement and the difficulty of holding a threshold accountable. Gillespie [2] documents how moderation became constitutive of platforms rather than incidental to them. Roberts [3] establishes the human labour behind moderation at scale and the conditions under which that labour is spent. Douek [4] argues for a shift from case-by-case, rights-absolutist reasoning toward proportionality and probability in speech governance, which is the normative counterpart of proportional enforcement as an engineering commitment. The present paper takes these framings as given and asks the next question: what does the routing table look like once a platform accepts that thresholds are value judgements?

Appeals, contestability, and procedural justice. Vaccaro, Sandvig, and Karahalios [5] show how users experience contesting algorithmic moderation decisions, and how the absence of an intelligible contest path corrodes perceived legitimacy independently of accuracy. The architecture described here treats the appeal path as a measurement instrument in addition to a fairness obligation, which is a stronger claim than the literature usually makes: appeals are the only channel through which a party with private knowledge and real stakes audits the system's calibration.

Contested ground truth and annotator disagreement. Aroyo and Welty [6] argue against the assumption that a single correct annotation exists for subjective tasks and propose treating disagreement as information. Davani, Díaz, and Prabhakaran [7] show empirically that majority-vote aggregation discards signal in subjective annotation. Both results are foundational to this paper's position that high inter-reviewer disagreement in a category is evidence about the policy, not about the reviewers, and that such categories should be escalated rather than trained on.

Deferral, selective prediction, and calibration. The idea that a model should abstain and route to a human on hard cases is well established. Wiener and El-Yaniv [8] formalise selective classification with a coverage-risk tradeoff; Narasimhan and colleagues [9] study estimators for learning to defer to an expert. Guo, Pleiss, Sun, and Weinberger [10] document that modern neural networks are poorly calibrated by default, which is directly relevant when band boundaries are derived from a confidence score. Settles [11] surveys active learning, whose central intuition, that labelling effort should be spent where the model is weakest, is the same intuition applied here to human review capacity rather than to annotation budget. The contribution of the present paper relative to this literature is that the deferral rule is not uncertainty alone. It is uncertainty conjoined with stakes, and the abstention region is sized by an external harm bound rather than by a coverage target.

Synthetic audio and AI-generated music detection. Detection of synthesised and manipulated audio has a mature benchmark tradition. Wang and colleagues [12] describe the ASVspoof 2019 database for synthesised, converted, and replayed speech. For music specifically, Cros Vila, Sturm, Casini, and Dalmazzo [13] examine the detection of AI-generated music and the adversarial dynamics of the task, which is the technical core of the detection layer described in Section 4. This paper does not contribute a detector or a benchmark; it treats detector quality as an input and asks what has to be true of the surrounding system for a detector to be usable in enforcement.

Streaming economics and coordinated fraud. Drott [14] analyses the economics of the streaming pool, including click fraud and fake-artist activity, which is the mechanism by which a false negative becomes a diffuse harm to real artists. Campaign-level detection has an established literature outside music: Pacheco and colleagues [15] present methods for uncovering coordinated networks, and Cresci [16] surveys a decade of social bot detection and the arms-race dynamic it produced. The cluster-review primitive in Section 5 is the operational counterpart of those findings inside an enforcement pipeline.

Scheduling, admission control, and prioritisation under a guarantee. Sections 5 and 6 are queueing arguments before they are anything else, and the closest engineering literature is cluster scheduling with differentiated service classes. Verma and colleagues [17] describe a production cluster manager in which latency-sensitive work is admitted against reserved quota while batch work runs on the remainder and is reclaimed when the latency-sensitive class needs its capacity back. That is structurally the same relation as the live path and the backfill in Section 6.3, and it is the reason the argument there is stated as an admission rule rather than left to a scheduler's fairness heuristics. The point of contact is also the point of difference. A general scheduler arbitrates between tenants of comparable standing, whereas the two workloads here differ in the economic cost of delay, so the priority relation is asymmetric by design rather than by configuration.

Industrial content identification. Large-scale audio fingerprinting is the closest existing industrial analogue to the detection layer of Section 4, and it faces recognisably the same operating constraints: a false-positive budget defined over identifiable rights holders, a reference set that is incomplete by construction, and a contest path attached to enforcement. Cano, Batlle, Kalker, and Haitsma [18] survey fingerprinting systems and the identification framework around them. The difference that matters here is that fingerprinting answers a matching question against a reference database, whereas AI-generation detection answers a question whose target is partly definitional, as Section 7.1 argues. A comparable false-positive discipline therefore has to be maintained without a comparable notion of a correct answer, which is what makes the layer separation of Section 7.2 load-bearing rather than tidy.

Prior work by the author. An earlier paper by the present author proposes a deterministic architecture for authenticating AI-generated content through provenance rather than through post hoc detection [19]. It is cited here as prior work only. It is not a contribution of the present paper, whose subject is the operation of a decision system rather than the design of a detector or a provenance scheme.

The gap. The literature treats detection as a modelling problem or governance as a normative problem, and rarely describes the join. What is missing is an account of the operational machinery that binds a precision

commitment to concrete artefacts: queue order, band width, backfill sequence, evidentiary standards, and an escalation path from reviewer disagreement to policy definition. That join is what this paper reports.

3. Problem Setting and Scale Constraints

3.1 Volume, arrival rate, and the deadline

The platform's catalogue exceeds 250 million tracks. Tens of thousands of new tracks arrive daily through distributor pipelines, each with audio and metadata. Every new upload requires a decision before it accrues meaningful streams, and the historical catalogue required a decision it had never received.

The daily arrival volume, while large, is bounded, which makes the live problem tractable. The catalogue is not bounded in the same practical sense: it is a fixed but enormous backlog that has to be traversed once. Section 6 treats that asymmetry as the central latency problem in the system.

The deadline is economic rather than technical. Royalties are a zero-sum pool, so a stream on an undetected AI-generated track is not a neutral error; it is a transfer away from real artists. Remediation after payout is expensive and incomplete. The system's objective is therefore not "score every track" but "score every track before money moves on it," which is a different objective with different consequences for prioritisation.

3.2 The two failure modes are not comparable

The most consequential design decision in the system was to stop treating precision and recall as two ends of one dial and start treating them as promises to two different constituencies.

Table 1. The asymmetry between the two failure modes, as it was used to make decisions. No incidence figures are given because none was measured or released; the table records the shape of the harms, which is what the thresholds were set against.

Dimension	False positive: a real artist flagged as AI	False negative: AI-generated content passes
Locus of harm	Concentrated on one identifiable person	Diffuse across the ecosystem
Who is hurt	A named artist: demonetisation, reputational damage, backlash, legal exposure	The royalty pool, recommendation quality, and the integrity of the catalogue
Visibility	Immediate, and it arrives with an advocate	Invisible until it is systemic; nobody escalates a royalty pool eroding one stream at a time
Recoverability	Largely unrecoverable: income and reputation, and the platform's legitimacy to enforce at all	Recoverable: recall can be improved next quarter
Institutional cost	Political permission to operate, with users, with the press, and internally	Coverage, and eventually an open door to industrial-scale fraud
How it was governed	A hard shipping gate, stated as a bound on artists wrongly flagged per month	A coverage and instrumentation obligation, monitored rather than gated

The consequence is a rule rather than a preference. The false-positive threshold was a hard shipping gate. The band boundaries described in Section 4 were derived from that gate, and when the derived human band was wider than review capacity liked, capacity flexed. Precision did not.

3.3 Harm as a count, not a rate

Presenting the same constraint to stakeholders in two different units produced two different conversations. Expressed as a rate, a false-positive bound reads as a technical parameter, and technical parameters are negotiable in roadmap discussions. Expressed as a count of artists wrongly flagged per month, it reads as what it is. Rates hide what thresholds decide.

Because the bound is load-bearing everywhere in this paper, it is worth defining exactly. Its unit is people over a period: the number of artists wrongly flagged per month that the programme committed in advance to stay within, held as a shipping gate rather than as a target to be approached. Its value is confidential and is not stated here, and no instrumented measurement of it is released, so the claim in Section 7.5 that the gate held as coverage grew is the author's account of how the constraint was governed rather than a result read off an evaluation.

The same translation was used to adjudicate the recall-for-precision tradeoff in Section 8. Both errors were converted into the same unit, people harmed rather than percentage points, before anyone was asked to choose between them. Rates make the tradeoff look technical. People make it look like what it is.

3.4 The strictest consumer sets the quality bar

Detection signals did not serve one purpose. They fed royalty decisions, editorial surfaces, and recommendations. Those consumers have different costs of error, and the platform cannot ship three different precisions from one signal without shipping the lowest one by accident.

The rule adopted was that the strictest consumer sets the quality bar: the downstream consumer with the highest cost of error defines the precision the whole platform ships at. This is expensive in the short run and it is the reason the signal was reusable at all. Building once to the highest precision requirement means every other consumer gets to build on top of it without renegotiating the bar.

3.5 Disclosure: what is quantified here

Two figures about the platform are given: a catalogue of more than 250 million tracks, and tens of thousands of new tracks ingested daily. Beyond those, no operational metric is released, and Section 10 states the omissions in full rather than repeating them here.

Some claims in the account do carry a magnitude, because the author described them that way: that most of the catalogue resolves in the cheap tier, that the backfill took months rather than weeks, that one reviewer decision could resolve hundreds of items. Those are reported as he described them and should be read as an operator's characterisation, not as a measurement made for this paper. Where a figure would ordinarily appear and none is available, the structural claim is made on its own and the absence is stated. The reason is simple: the operational data is confidential, and a fabricated decimal in this domain is the most checkable claim in the paper and the least defensible.

4. System Architecture

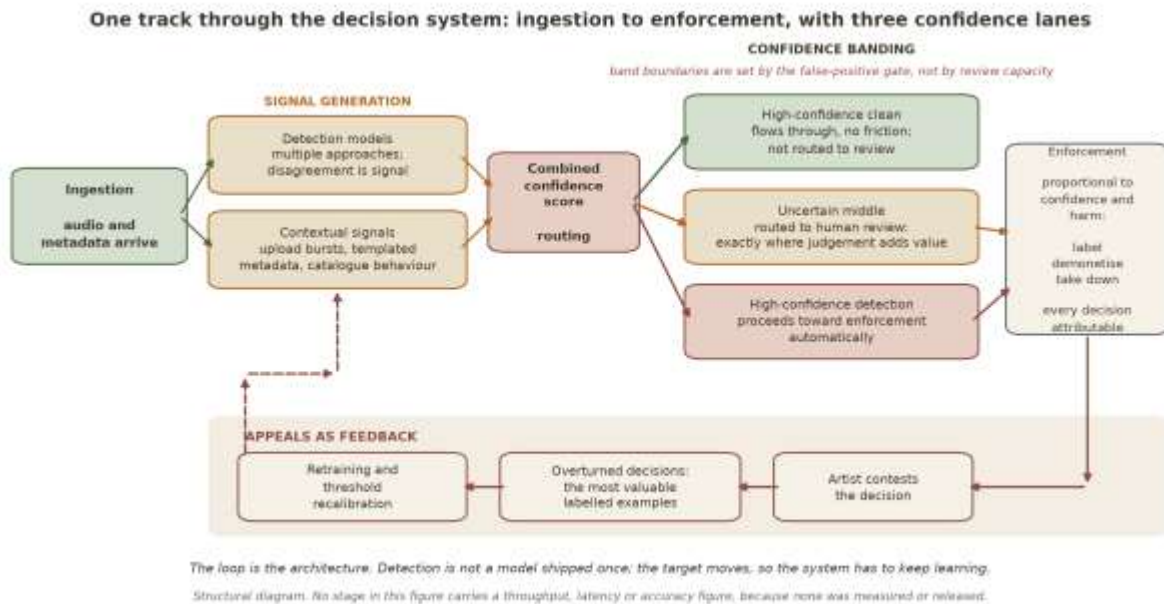


Figure 1. One track through the decision system, from ingestion to enforcement, with the three confidence lanes and the appeal feedback path. Structural diagram: no stage carries a throughput, latency, or accuracy figure, because none was measured or released.

Figure 1 traces a single track. The stages are described below in the order the track meets them.

4.1 Ingestion

A track arrives through a distributor pipeline with its audio and its metadata. Before it is eligible for anything downstream, it enters the detection flow. The design intent is explicit in the ordering: the goal is a decision before the track accrues meaningful streams, not after. Detection placed after distribution is a reconciliation process, not an enforcement one.

4.2 Signal generation

Two families of signal are produced.

The first is audio analysis. The audio runs through detection models that score the likelihood that it is AI-generated. More than one detection approach was used deliberately, rather than betting the system on a single model. Different techniques fail differently, and that property is useful twice: it reduces correlated failure, and it makes disagreement between approaches an informative signal in its own right rather than noise to be averaged away. Section 7 develops disagreement into a standing measurement instrument.

The second is contextual. Upload patterns, metadata consistency, and catalogue behaviour are examined alongside the audio. This is where the system stops being a classifier and starts being an integrity system. A single suspicious track means little. A burst of uploads sharing templated metadata and a common catalogue signature means a great deal, and it means it even when each individual track's audio score is unremarkable. The contextual layer is also what makes the cluster-review primitive of Section 5 possible, because the same linkage evidence that raises suspicion is the linkage evidence that lets one human decision resolve a whole campaign.

4.3 Scoring and confidence banding

The signals combine into a confidence score, and the score routes the track into one of three lanes.

Table 2. The three confidence lanes. The boundaries between lanes are owned by the false-positive gate, not by review capacity.

Lane	Routing	Action	Human involvement
High-confidence clean	Flows through with no friction	None; the track is eligible downstream	None; the confident clean traffic is never routed to a reviewer
Uncertain middle	Routed to human review, ordered by stakes	Decided by a reviewer in stakes order, or aggregated into a cluster case before a reviewer sees it; a low-stakes item can wait and be re-scored later	This is the only lane where human judgement is spent on the score itself
High-confidence detection	Proceeds toward enforcement	Proportional to confidence and harm, from labelling and demonetisation up to takedown	The score alone does not create a review queue entry; every action is attributable, recording which signals fired, at what threshold, and whether a model or a human decided it

This is the core scaling move, and it is worth stating precisely what it does and does not accomplish. It does not make the model better. It removes the overwhelming majority of traffic from human consideration, in both directions, so that the human capacity that exists is spent on the slice where judgement changes the outcome. That slice, plus the appeals described in Section 5.4, is what humans see. Routing confident traffic to humans does not add judgement; it adds latency and cost, and it spends capacity the uncertain slice needs.

The ownership of the band boundaries is the part teams most often get backwards. The boundaries are set by the false-positive gate, not by review capacity. If the band those boundaries define is wider than review capacity would like, the human band is widened before precision is loosened. That ordering is the whole of the commitment, and it is worth being clear about what kind of thing it is. Nothing in the machinery prevents a threshold from being moved, so the ordering is a governance rule rather than a property of the architecture, and it holds for exactly as long as the team that owns the gate keeps holding it.

4.4 Proportional and attributable enforcement

Action is proportional to confidence and to harm. The available actions form a ladder rather than a switch: labelling, demonetisation, and takedown are different responses, and the system is designed so that a moderate-confidence signal on a low-harm case does not have to reach for the top of the ladder in order to do something. Every enforcement decision is attributable. The record carries which signals fired, at what threshold, and whether the decision was made by a model or by a human. This is not bureaucracy and it is not primarily a compliance artefact. It is the precondition for an honest appeal: a decision whose basis cannot be reconstructed cannot be meaningfully contested, and an appeal process that cannot inspect the basis of the decision it is reviewing is a customer-service ritual rather than an audit.

The asymmetry of Section 3.2 reappears here as an evidentiary rule. Enforcement actions required corroborating signals beyond the model score alone. A clean pass did not. Symmetric certainty is not required when the costs of error are not symmetric.

4.5 Appeals as feedback

An artist can contest a decision. Appeal outcomes flow back into the system as labelled data, and overturned decisions are the most valuable examples the system produces, because an overturn marks precisely the region where confidence was miscalibrated. Section 7.4 treats appeals as a measurement instrument and Section 5.4 treats them as a queue-priority rule; here the point is architectural. The loop is the architecture. Detection is not a model that gets shipped once. It is a system that has to keep learning, because the other side keeps learning.

4.6 What the architecture is optimising

It is worth naming the objective the architecture serves, because it is not accuracy. The system is optimised so that the small number of decisions that matter most, the ones where a real person can be harmed in a way that cannot be undone, receive disproportionate attention, and so that miscalibration surfaces quickly and in analysable form. Every structural choice above serves one of those two ends.

5. Review Allocation and the Uncertain Band

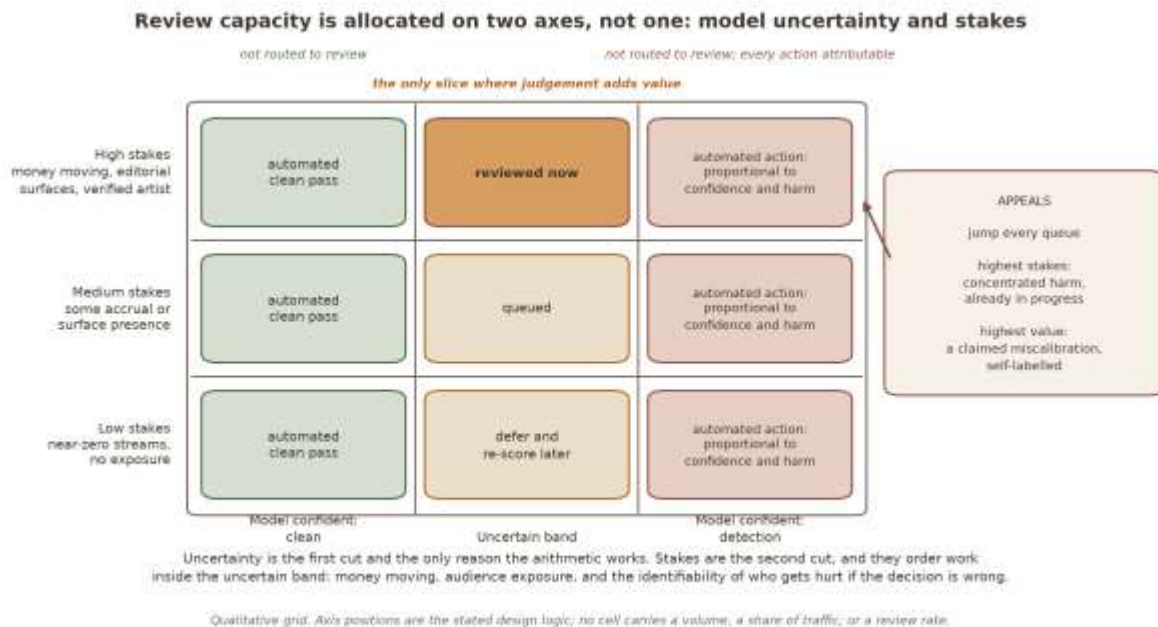


Figure 2. Review capacity is allocated on two axes, model uncertainty and stakes, with appeals as a standing exception that jumps every queue. Qualitative grid: axis positions record the design logic; no cell carries a volume, a share of traffic, or a review rate.

5.1 Uncertainty alone is the wrong triage key

The naive allocation rule is "send humans the borderline cases." It is the right first cut and the wrong complete answer.

It is the right first cut because it is the only reason the arithmetic works. Humans never see the traffic the models are confident about, in either direction. Nothing else about the system matters if that filter is not in place.

It is the wrong complete answer because within the uncertain band the items are not interchangeable. A borderline track with near-zero streams and no surface presence can wait, or can simply be re-scored later. A borderline track that is accruing revenue, sitting on editorial surfaces, or attached to a verified artist's profile needs a decision now. Treating those two items as equivalent because their confidence scores are equivalent is a category error about what the queue is for.

5.2 Stakes as the second axis

Stakes were a function of three things: whether money is actively moving, how much audience exposure the item has, and how identifiable the party is who gets hurt if the decision is wrong.

The third component is the one that carries the asymmetry from Section 3.2 into the queue. A wrongful flag on a real artist is the most expensive mistake the system can make, so anything touching a verified artist received priority. That is the false-positive gate expressed as queue order rather than as a threshold. The same

commitment appears in two places in two different forms, which is a property worth designing for: a constraint that only exists in one artefact will eventually be traded away in the artefact it does not appear in.

The low end of the stakes axis does the same work in reverse. A borderline track with near-zero streams can wait, or can simply be re-scored later, so spending a reviewer on it now buys nothing that waiting does not buy more cheaply. Stakes ordering is what makes that difference visible in the queue instead of leaving it hidden inside an arrival order.

5.3 Cluster review: per campaign, not per item

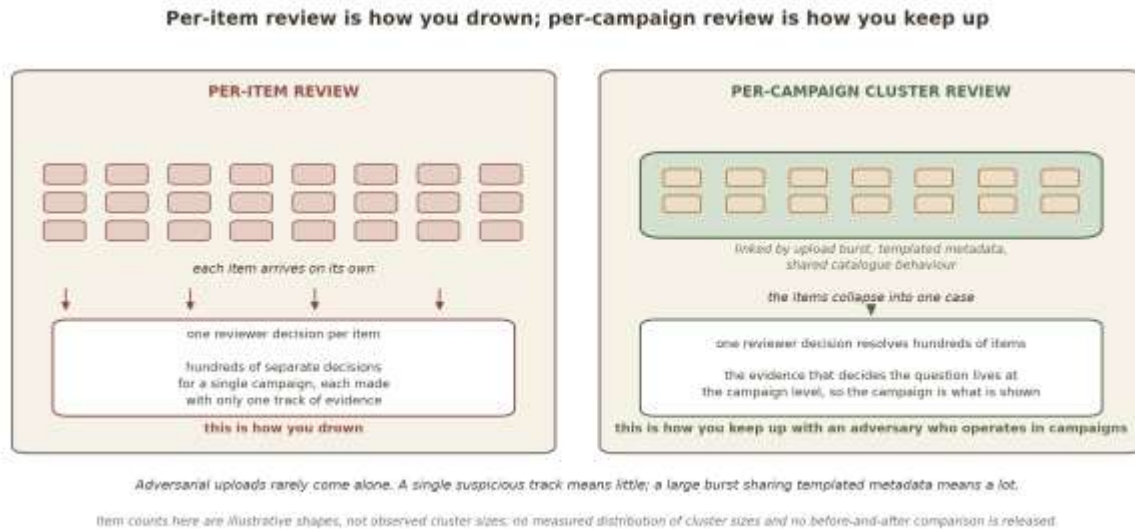


Figure 3. Per-item review versus per-campaign cluster review. The item counts drawn here are illustrative shapes rather than observed cluster sizes. No measured distribution of cluster sizes is released, and no before-and-after comparison of the two regimes is reported.

Adversarial uploads rarely arrive alone. When contextual signals tied tracks together, by shared upload burst, templated metadata, or common catalogue behaviour, the linked items were collapsed into a single case before a reviewer saw them. One reviewer decision could then resolve hundreds of items rather than one.

Per-item review is how you drown. Per-campaign review is how you keep up with an adversary who operates in campaigns.

The throughput arithmetic is the obvious half of the benefit. The less obvious half is that aggregation changes what the reviewer is looking at, and changes it for the better. The contextual signals of Section 4.2 make the point on their own: a single suspicious track means little, while a large burst of uploads sharing templated metadata means a great deal. The evidence that decides the question exists at the campaign level, so presenting the campaign is presenting the evidence, and presenting one track at a time is withholding most of it.

5.4 Appeals jump the queue

Appeals are a standing exception to the stakes ordering. An appeal means a real person is asserting that the system got it wrong about them specifically. That makes it simultaneously the highest-stakes review in the system, because the harm is concentrated and already in progress, and the highest-value training data in the system, because it is a claimed miscalibration that arrives self-labelled by the party with private knowledge.

Deprioritising appeals to protect throughput is optimising the metric that matters least.

5.5 The human queue as a dashboard of model debt

The honest constraint underneath all of the above is that review capacity is a budget, and the budget is spent where the model is weakest today. That reframing turns the queue into an instrument.

The shape of the human queue was treated as a dashboard of model debt. If one category kept flooding the uncertain band, that was not a staffing problem. It was the next modelling investment, announcing itself. If reviewers within a category disagreed with each other rather than with the model, that was not a training-data problem either, and Section 7.4 explains why it goes to policy instead. Read that way, the queue is not only a workload to be staffed. It is the system's own report on where its models are weakest, which is the same document as the roadmap.

The goal was never to review more. It was to need less review for the same precision. Those are different objectives and they produce different roadmaps.

6. Backfill Under a Live SLA

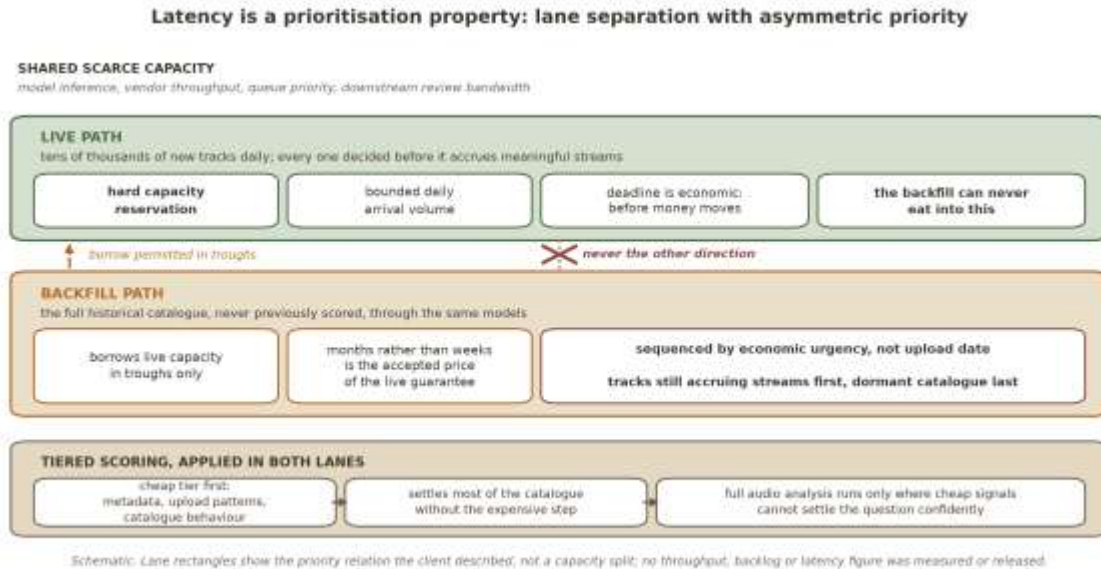


Figure 4. Lane separation with asymmetric priority, plus tiered scoring applied in both lanes. Schematic: the rectangles show the priority relation, not a capacity split. No throughput, backlog, or latency figure was measured or released.

6.1 The problem was not making one decision fast

The hardest engineering obstacle in the system was not decision latency in isolation. It was making decisions fast while re-deciding the past.

The live pipeline has a natural deadline and a bounded input, so it is tractable. The problem arrived with the commitment to backfill detection across the historical catalogue: tracks that had never been scored, running through the same models, without degrading the latency of the live path.

The two workloads compete for everything that is scarce: model inference capacity, vendor throughput, queue priority, and review bandwidth downstream. Naive interleaving produces the worst available outcome, in which the freshest and most economically urgent decisions wait behind a multi-month batch job.

Three mechanisms resolved it.

6.2 Tiered scoring

Full audio analysis is the expensive step, so not every track earns it. Cheap signals run first: metadata, upload patterns, catalogue behaviour. The expensive models run only where the cheap signals cannot already settle the question confidently. Most of the catalogue resolves in the cheap tier.

This is the same principle as the human review band, one level down the stack. The latency budget goes where the uncertainty is. Both mechanisms are instances of a single rule: spend the expensive resource, whether that is a human or a model, only where the cheap resource has failed to produce confidence.

6.3 Lane separation with asymmetric priority

The live path received a hard capacity reservation that the backfill could never eat into, no matter how far behind the backfill ran. The relation was deliberately asymmetric: the backfill borrowed live capacity in troughs, but the live path never borrowed from the backfill's floor, because the backfill had no floor to defend.

The consequence was accepted explicitly rather than discovered. The backfill would take months rather than weeks, and that was the price of the live guarantee. It was the right price, and the reason is economic rather than technical. A new track scored late costs money now. A track uploaded years ago and scored a week later costs almost nothing more than a week earlier.

Stating the asymmetry as a rule, rather than leaving it to a scheduler's fairness heuristics, is what made it hold under pressure. A fair scheduler is exactly the wrong instrument here, because the two workloads do not deserve fairness.

6.4 Sequencing the backfill by economic urgency

Chronological order is the obvious sequence for a backfill and the wrong one. The backfill was ordered by where being wrong in either direction was most expensive: tracks still accruing streams first, dormant catalogue last. That reordering changed which question the programme was accountable for. "When will the backfill finish?" became a trailing detail. The metric that mattered was how quickly unscored revenue was going to zero. Those two questions have different answers and only one of them is about harm.

6.5 Latency is a prioritisation property

The general lesson from this part of the system is the one that generalises furthest beyond music. At this volume, latency is not an infrastructure property. It is a prioritisation property. The models were never made much faster. The team got ruthless about which items deserved fast, and that turned out to be the larger lever by a wide margin.

7. Measuring Correctness Without an Oracle

The model measures; the policy decides. Three layers, and four independent ways miscalibration becomes visible

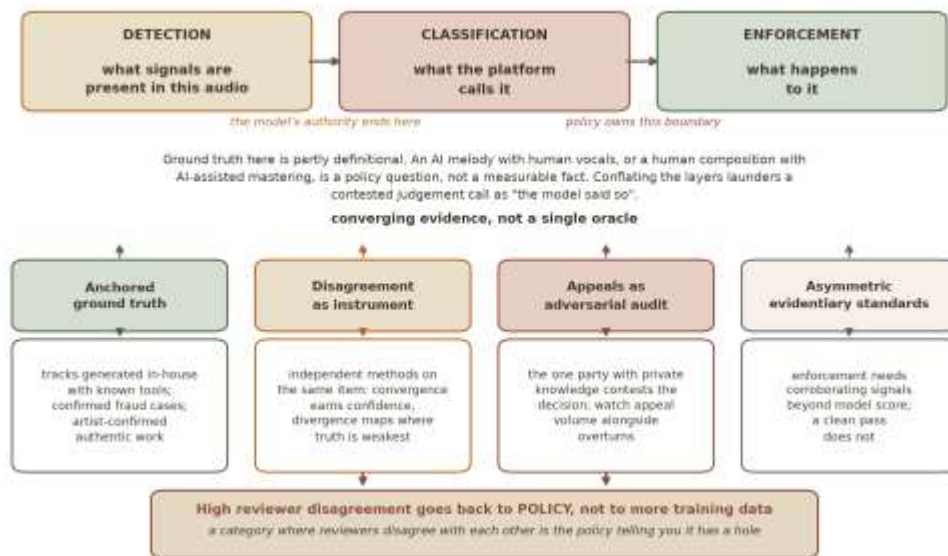


Figure 5. The separation of detection, classification, and enforcement, and the four independent instruments through which miscalibration becomes visible. High reviewer disagreement returns to policy rather than to training.

7.1 Ground truth is partly definitional

The uncomfortable admission belongs at the front. For AI-generated content, ground truth is not merely hard to obtain. It is partly a matter of definition.

A fully synthetic track is one thing. A human-composed track with AI-assisted mastering is another. An AI melody with human vocals is a third. Whether any of those is "AI music" is a policy decision, not a measurable fact. No amount of model improvement resolves it, because the disagreement is not about the audio.

7.2 The model measures; the policy decides

The first move, therefore, is refusing to let the model answer a question the policy has not answered. Three layers were kept separate:

- Detection: what signals are present in this audio.
- Classification: what the platform calls it.
- Enforcement: what happens to it.

The model's authority ends at the first layer. Conflating the layers is how systems end up laundering contested judgement calls as "the model said so," and it has a second cost that is easy to miss: once the layers are fused, a change in policy requires retraining the measurement. Keeping them separate means a policy answer can change without touching the detector, which is the property that makes a contested domain operable at all.

7.3 Converging evidence rather than a single oracle

Given a partly definitional target, correctness was estimated through four instruments with independent failure modes.

Anchored ground truth where it exists. Some truth is purchasable: tracks generated in-house with known tools, confirmed fraud cases from investigations, artist-confirmed authentic work. This set is small relative to the catalogue, and its role is not to measure system accuracy. Its role is to anchor calibration. A model that fails on cases where truth is known cannot be trusted on cases where it is not.

Disagreement as instrument. Multiple detection approaches with different failure modes, plus human reviewers, score the same items. Where independent methods converge, confidence is earned. Where they diverge, the divergence is not noise to average away; it is a map of exactly where ground truth is weakest. Inter-rater disagreement among human reviewers was read the same way.

Appeals as adversarial audit. An appeal is the one measurement in the system where a party with real stakes and private knowledge challenges the decision, and in this domain the artist knows how the track was made. Overturn outcomes, sliced by category and by confidence band, were the closest thing available to a field test of correctness. One caution was treated as standing: a low overturn rate can mean the system is accurate or it can mean the appeal process is too burdensome to use, so appeal volume was watched alongside overturn outcomes. Silence is not agreement.

Asymmetric evidentiary standards. Because the harms are asymmetric, the burden of proof was too. Enforcement required corroborating signals beyond the model score. A clean pass did not.

7.4 Disagreement routes to policy, not to training

The rule that did the most work in this part of the system is also the easiest to get wrong. High-disagreement categories, where reviewers ruled inconsistently with each other, were not treated as model weakness and were not fed back as training data. They were escalated to policy.

The reasoning is that such a category is not reviewer failure. It is the policy reporting that it has an undefined region, and the destination for an undefined region is the body that defines it rather than the training set. Retraining on those decisions would instead launder a value judgement through a model confidence score, which is the failure described in Section 9.3, arriving through the training data rather than through the threshold. Neither a threshold change nor more labelling resolves a rule the institution has not written.

7.5 What could and could not be claimed

Coverage is the cleanest claim the programme can make. The system moved from partial visibility to detection across the full catalogue, including the historical backfill, with every new upload scored before it accrues meaningful streams. Coverage is the foundation metric, because accuracy over whatever sample happened to be scored is a vanity number.

The result of which the author is proudest is that precision held as coverage scaled: the gate set at the start was the gate the system shipped at full catalogue scale, and it held while scored volume grew by orders of magnitude, which required repeated recalibration as new generation tools appeared. Section 3.3 states what that claim is and is not: the author's account of how the constraint was governed rather than a result read off an evaluation. This paper cannot say which of the four instruments in Section 7.3 the bound was checked against, or how often, and the mechanism was the governance ordering of Section 4.3 rather than a property of the machinery.

Reviewer leverage and latency improved structurally rather than measurably. Routing confined human review to the uncertain band, so review demand grew far more slowly than catalogue volume; cluster review meant one decision could resolve an upload campaign; every new track was decided at ingestion rather than after accrual; and stakes-based triage spent the queue's latency on purpose. Beyond the characterisation in Section 5.3, no distribution of items resolved per reviewer decision, no before-and-after comparison of the two review regimes, no share of scored traffic that reached a human, and no time-to-decision figure is released. On appeals, what can be claimed is the loop rather than a curve: outcomes fed back as labelled data and as threshold calibration, and the overturn slices of Section 7.3 were consumed as calibration evidence, but whether the overturn rate held steady or declined is not stated because it was not made available for publication.

Accuracy in the absolute sense is unmeasured, and the paper says so plainly. Ground truth here is partly definitional, so no oracle accuracy number exists, and there was never a correctness number the author would defend as absolute. What existed instead was a system in which miscalibration had several independent ways of becoming visible, and a discipline of treating contested cases as inputs to policy rather than as errors in the model. When ground truth is subjective, the measurable thing is not whether the system is right. It is how quickly it finds out when it is wrong.

8. Deliberate Tradeoffs

Three sacrifices were made on purpose. Naming them is part of the contribution, because the alternative, a paper in which the architecture has no costs, is not an account of a real system.

Table 3. The three deliberate sacrifices and the irreversibility comparison that decided each one.

Sacrifice	What was given up	Why it was acceptable	The irreversible thing that was protected
Recall, to protect precision	Accepting in writing that AI-generated content gets through, more than would be caught at a looser threshold	The two harms are not symmetric, and trust compounds asymmetrically; both errors were converted into people harmed before anyone chose	An artist's income and reputation, and the platform's legitimacy to run an enforcement programme at all
Building, in favour of buying, and speed in favour of control	Some detection capabilities were sourced externally; the author argued to build one of them and lost	What survived the loss mattered more than the decision: the evaluation framework built to make the case became how every external capability was held accountable, against internal data, thresholds, and failure modes	The definition of good. Vendors can supply signals; they cannot own the precision bar
Other roadmaps	Detection was sequenced ahead of work an adjacent team was counting on, at a real cost to their timeline	The scarcest resource at a given moment is not engineering capacity; it is the window before a threat becomes systemic. Detection had a closing window; the deferred work did not	Only defensible if said out loud, with a date on the deferred work, and the date kept

8.1 Recall for precision

Holding a hard false-positive gate means accepting, in writing, that AI-generated content gets through, and more of it than would get through at a looser threshold. That is uncomfortable to defend, because misses are the failure mode critics can point to and the failure mode that generates headlines.

The tradeoff was acceptable because the two harms are not symmetric and because trust compounds asymmetrically as well. Recall can be improved next quarter. A wave of wrongful takedowns is not something a platform easily recovers from, and it does not only cost the affected artists. It costs the political permission on which the entire enforcement programme runs.

The mechanism that made the decision reviewable was the unit conversion from Section 3.3. Both errors were translated into people harmed rather than percentage points before anyone was asked to choose.

8.2 Build versus buy, and losing the argument

Some detection capabilities were sourced externally rather than built in-house. The author argued for building in one of those decisions and lost.

What survived the loss mattered more than the decision. The evaluation framework that had been built to make the case became the mechanism for holding every external capability accountable: measured against internal data, internal thresholds, and internal failure modes rather than against a vendor's benchmark. The tradeoff was acceptable because the thing that cannot be outsourced was kept. Vendors can supply signals. They cannot own the precision bar.

This generalises into a practical rule for build-versus-buy decisions in trust and safety specifically. The question is not whether an external capability is good. It is whether the organisation retains the ability to define and measure good independently of the supplier; because the supplier's failure modes are not the organisation's failure modes and the supplier's threshold is not the organisation's harm bound.

8.3 Other roadmaps

Scaling detection meant sequencing it ahead of work other teams were counting on. That was a forward-first call with a real cost to an adjacent team's timeline, and the cost was owned directly to the affected team rather than left to be discovered.

The deciding logic was that at a given moment the scarcest resource is not engineering capacity but the window before a threat becomes systemic. Detection had a closing window. The deferred work did not. That is a

defensible tradeoff only under three conditions: it is said out loud, the deferred work gets a date, and the date is kept.

8.4 The meta-rule: compare irreversibility

Every tradeoff in the system was forced into the same comparison, which is a comparison of irreversibility rather than of magnitude.

Recall can be recovered. A deferred roadmap can be rescheduled. Wrongfully flagged artists and burned trust mostly cannot be recovered. When the team was unsure, it sacrificed the recoverable thing.

The reason to name this as a rule rather than as an instinct is that magnitude comparisons are the default in engineering organisations, and magnitude comparisons systematically favour the legible error. Irreversibility comparisons do not, and they are the only comparison that gets the asymmetry of Section 3.2 into the decision at the moment the decision is made rather than in the retrospective afterwards.

9. What Teams Underestimate

This section generalises. Five things are consistently underestimated by teams building trust and safety systems, listed in roughly the order teams learn them the hard way. Each one is a consequence of something earlier in this paper.

9.1 The adversary's iteration speed, and their own

Teams plan detection like a feature: ship, measure, improve on a quarterly cadence. The other side ships weekly, has no review process, and treats enforcement actions as free quality assurance on their evasion. The moment a platform enforces, it has taught the adversary something.

Systems built as one-time models decay silently, which is worse than decaying loudly. The build has to assume the target moves. That is why the feedback loop in Section 4.5 is described as the architecture rather than as an enhancement to it. This dynamic is well documented outside music, in the arms race between social bot detection and evasion [16], and in the adversarial dynamics reported for AI-generated music detection specifically [13].

9.2 False positives as an existential risk rather than a metric

Every team says precision matters. Few price in what one wave of wrongful enforcement does to the legitimacy of an entire programme.

Recall failures cost coverage. Precision failures cost the political permission to operate: with users, with the press, and inside the company. In practice, the small minority of decisions that were wrong consumed more organisational attention than the large majority that were right, and that imbalance is not a dysfunction to be managed away. It is the actual cost structure of the domain, and a plan that does not budget for it is not a plan.

9.3 Policy debt

Teams assume the hard part is detection and the policy is settled. It is usually the reverse. The model surfaces cases the policy never contemplated, and every undefined region becomes either inconsistent enforcement or reviewer improvisation.

High reviewer disagreement is not a training problem. It is the policy announcing that it has a hole. Teams that do not route those cases back to policy end up laundering value judgements through model confidence scores, which is the failure mode Section 7 is designed to prevent. This is also where the annotation literature and the operational reality meet: the finding that disagreement in subjective tasks carries signal rather than noise [6], [7] is, in an enforcement system, a statement about institutional debt rather than about labelling technique.

9.4 The appeal path

Appeals get built last, staffed least, and treated as overhead. They are actually three things at once: the highest-stakes queue in the system, the only adversarial audit of correctness available, and the cheapest labelled data the organisation will ever get.

A trust system without a serious appeal loop is not measuring itself. It is grading its own homework. The literature on contestability reaches the same conclusion from the user's side [5]; the argument here is that the engineering case is just as strong as the fairness case, and that a team that cannot win the fairness argument internally can often win the calibration one.

9.5 The diffuse failure mode has no complainant

The last underestimation is structural rather than technical, and it is the one that most reliably survives good intentions.

Concentrated harms have advocates. The wrongly flagged artist emails the chief executive. Diffuse harms have nobody: no individual escalates a royalty pool eroding one stream at a time. Organisations therefore over-rotate toward the legible error and starve the invisible one, right up to the point where the invisible one is systemic. The fix is not vigilance, because vigilance is a property of attention and attention follows escalation. The fix is instrumentation. The dashboard for the harm that has no complainant has to be built deliberately, because no ticket queue will ever produce it.

10. Limitations

This paper's limitations are substantial and specific, and they bound what a reader should take from it.

Single-platform experience. Everything reported here comes from operating one system at one large-scale music streaming platform. The architecture was shaped by that platform's economics, in particular a zero-sum royalty pool that gives the decision an unusually crisp financial deadline. Platforms whose enforcement decisions do not sit directly upstream of a payout may find the prioritisation arguments in Section 6 less binding, even where the routing arguments in Sections 4 and 5 transfer.

One content domain. The content is music audio, and the harms are economic and reputational. Domains where the harm is physical safety, where the content is itself traumatic, or where the affected party is not an identifiable rights holder have different asymmetries, and the routing arguments here should be re-derived rather than transported. This paper makes no claim about reviewer welfare design, which is a substantial literature in its own right and is outside what the experience reported here can speak to.

No controlled evaluation. This is an architecture and governance experience report. No experiment was run for this paper. There is no ablation showing what two-axis allocation contributes over uncertainty-only allocation, no measured comparison of cluster review against per-item review, and no controlled test of the backfill sequencing rule against chronological order. The claims are design rationale and observed operational behaviour, not measured effects.

No released operational metrics. The underlying operational data is confidential. Consequently this paper states no automated-versus-reviewed share, no peak-throughput figure, no appeal-overturn rate or trend, no measured distribution of items resolved per reviewer decision and no before-and-after comparison of per-item against cluster review, no recalibration cadence, no count of downstream consumers of the detection signal, and no time-to-decision figure. Two quantities about the platform are given: a catalogue exceeding 250 million tracks, and tens of thousands of new tracks ingested daily. A small number of further magnitudes appear as the author characterised them, and Section 3.5 marks them as characterisations rather than measurements. Where a structural claim in this paper would normally carry a supporting number, the claim stands alone and the gap is marked. A reader is entitled to treat the unquantified claims as weaker for that reason; the alternative, supplying an estimate, would make them weaker still and less honest.

Confidentiality also bounds the failure narrative. The divergence patterns and category discoveries referenced in Section 7 are described structurally rather than with the specific categories, tools, and vendors involved. Some of the most instructive detail in an operational account of this kind is exactly the detail that cannot be published.

No formal model. The paper offers no formal treatment of the two-axis allocation rule, although the selective prediction and deferral literature [8], [9] provides the natural vocabulary for one. Formalising the rule with an explicit harm-bound constraint, rather than a coverage target, is the obvious next piece of work and it does not require access to confidential data.

11. Conclusion

At a catalogue of more than 250 million tracks, with tens of thousands more arriving daily, human review is not a policy option and the machine is not a convenience. That single arithmetic fact reorganises everything: what a threshold is, who owns it, which decisions may be automated, and what the humans in the system are for.

The account in this paper reduces to six commitments. Treat the two failure modes as promises to different constituencies, and express the false-positive bound as a count of people rather than as a rate, because rates hide what thresholds decide. Route the confident majority away from humans in both directions, and spend the resulting capacity on the uncertain band ordered by stakes rather than by arrival. Review campaigns, not items, because the adversary operates in campaigns and because per-item queues destroy the judgement they depend on. Separate the live path from the historical backfill with a hard reservation, and sequence the backfill by economic urgency, because at this volume latency is a prioritisation property and not an infrastructure one. Keep detection, classification, and enforcement in separate layers, so the model measures and the policy decides. And build several independent ways for miscalibration to become visible, then treat the contested cases as inputs to policy rather than as errors in the model.

The deciding rule behind all six is a comparison of irreversibility rather than of magnitude. Recall can be recovered. A deferred roadmap can be rescheduled. Wrongfully flagged artists and burned trust mostly cannot. When the answer was unclear, the recoverable thing was sacrificed.

The claim the author would most like to leave with practitioners is the one that took longest to learn. Teams build trust and safety as if the problem is classification. The problem is governance under adversarial pressure, at a scale where every threshold is a values statement. The classifier is the easy part.

References

- [1] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, 2020. doi:10.1177/2053951719897945
- [2] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press, 2019. doi:10.12987/9780300235029
- [3] S. T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press, 2019. doi:10.12987/9780300245318
- [4] E. Douek, "Governing Online Speech: From 'Posts-As-Trump' to Proportionality and Probability," working paper, 2020. doi:10.2139/ssrn.3679607
- [5] K. Vaccaro, C. Sandvig, and K. Karahalios, "'At the End of the Day Facebook Does What It Wants': How Users Experience Contesting Algorithmic Content Moderation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1-22, 2020. doi:10.1145/3415238
- [6] L. Aroyo and C. Welty, "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation," *AI Magazine*, vol. 36, no. 1, pp. 15-24, 2015. doi:10.1609/aimag.v36i1.2564
- [7] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92-110, 2022. doi:10.1162/tacl_a_00449
- [8] Y. Wiener and R. El-Yaniv, "Agnostic Pointwise-Competitive Selective Classification," *Journal of Artificial Intelligence Research*, vol. 52, pp. 171-201, 2015. doi:10.1613/jair.4439
- [9] H. Narasimhan, W. Jitkrittum, A. K. Menon, A. S. Rawat, and S. Kumar, "Post-Hoc Estimators for Learning to Defer to an Expert," in *Advances in Neural Information Processing Systems 35*, 2022, pp. 29292-29304. doi:10.52202/068431-2124
- [10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR vol. 70, 2017, pp. 1321-1330.
- [11] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer, 2012. doi:10.1007/978-3-031-01560-1
- [12] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, art. 101114, 2020. doi:10.1016/j.csl.2020.101114
- [13] L. Cros Vila, B. L. T. Sturm, L. Casini, and D. Dalmazzo, "The AI Music Arms Race: On the Detection of AI-Generated Music," *Transactions of the International Society for Music Information Retrieval*, vol. 8, no. 1, 2025. doi:10.5334/tismir.254
- [14] E. Drott, *Streaming Music, Streaming Capital*. Durham, NC: Duke University Press, 2023. doi:10.1215/9781478027874
- [15] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 455-466, 2021. doi:10.1609/icwsm.v15i1.18075
- [16] S. Cresci, "A decade of social bot detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72-83, 2020. doi:10.1145/3409116
- [17] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at Google with Borg," in *Proceedings of the Tenth European Conference on Computer Systems*, 2015, pp. 1-17. doi:10.1145/2741948.2741964
- [18] P. Cano, E. Battle, T. Kalker, and J. Haitsma, "A Review of Audio Fingerprinting," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 41, no. 3, pp. 271-284, 2005. doi:10.1007/s11265-005-4151-3
- [19] B. Kandati, "From Detection to Provenance: A Deterministic Architecture for Authenticating AI Generated Content," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 7, no. 1, pp. 238-240, 2026. doi:10.63282/3050-9262.ijaidsm-l-v7i1p139